DOCUMENT RESUME

ED 441 032 TM 030 831

AUTHOR Bashook, Philip
TITLE Assessing Clinical Judgment Using Standardized Oral Examinations.
PUB DATE 2000-04-00
NOTE 7p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Higher Education; Medical Education; *Medical Students; *Standardized Tests; *Test Construction
IDENTIFIERS *Clinical Competence; *Oral Examinations

ABSTRACT

This paper describes the use of oral examinations to assess the clinical judgment of aspiring physicians. Oral examinations have been used in U.S. medicine since 1917. Currently, 15 member boards of the American Board of Medical Specialties administer 17 different standardized oral examinations to approximately 10,000 physician candidates annually. The oral examination used in specialty certifying examinations is a carefully crafted series of standardized examination sessions or stations similar to a role-play situation. These examinations rely on standardization of examiners and standardization of cases. Several different score approaches are used, but in principle the expectation is to generate as many separable scores as feasible from as many cases as possible. In some, but not all, examinations the cases and examiners are calibrated using item response theory methods. The standardized oral examination is one potential way to measure the clinical judgment of professionals. (Contains 15 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 032

Assessing Clinical Judgment Using Standardized Oral Examinations

Philip Bashook

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

P. Bashook

AERA

TM030831

BEST COPY AVAILABLE

Assessing Clinical Judgment using Standardized Oral Examinations

Philip Bashook, EdD
American Board of Medical Specialties (ABMS)

Paper presented at the
American Educational Research Association
Annual Meeting
April 24-28, 2000
New Orleans, LA

Oral examinations have been used in American medicine to assess clinical judgment since 1917 when the first specialty certifying board was established. Currently, fifteen Member Boards of the American Board of Medical Specialties administer annually 17 different standardized oral examinations to approximately 10,000 physician candidates.[1] Successfully passing the oral examination is the final step in initial certification. Physicians must qualify for entry to the oral exam by completing three to eight years of approved post medical school residency training, demonstrating appropriate professional behavior including meeting an acceptable level of performance during the residency, posses a valid medical license in the US or Canada, and pass a standardized written examination of knowledge.[2]

The medical and surgical certifying oral exams are intended to assess clinical judgment or the application of what Bordage refers to as "elaborated knowledge."[3] The abilities defined as clinical judgment include: clinical reasoning, application of knowledge, and knowing one's own limits. Generalizing to other professions these judgment abilities can be characterized as follows:

1. Rapidly identify, interpret, and synthesize key findings when presented with a realistic professional problem.
2. Use knowledge effectively and efficiently to make decisions about defining the problem and solving it.
3. Demonstrate recognition of personal limits in knowledge and expertise appropriate to the level of expertise expected for certification in the profession.

Nearly all the oral exams measure the first two abilities directly as "patient work-up," diagnosis or differential diagnosis," and "treatment plans or patient management." Other attributes reported as measured in some oral exams include "professionalism" (in seven exams), and "interpersonal skills" (in six exams). The third ability of clinical judgment, "know own limits," refers to recognizing the limits of current scientific knowledge when applied to a specific patient situation, and recognizing one's own limits in knowledge and abilities.

Description of the standardized oral examination

Unlike the oral defense of a PhD dissertation an oral examination when standardized is not a general question and answer session between the candidate and a panel of experts.[4] The oral exam used in specialty certifying exams is a carefully crafted series of standardized examination sessions or stations similar to a "role-play simulation." Realistic patient cases are the focus of discussion and the physician examiner serves as the simulated patient database, the questioner, and evaluator. The examiner begins the oral exam by providing the case stimulus as a brief scenario with or without visual aids. The examiner provides the findings relevant to the case as

requested by the candidate as well as questioning the candidate about the reasons and rational for the response. Memorized textbook-based answers are not sufficient responses to questions. The examiner uses probing questions sometimes involving positing variations on the original case to verify that the candidate can apply current scientific knowledge or reason at the appropriate level of expertise.

A typical oral examination lasts 2.5 hours with a range of one hour to 3.5 hours. Most exams contain three to six stations with one examiner per station. Three exams use two or three examiners per station but fewer stations. Between 6 and 10 cases can be evaluated in 30 minutes. A number of anecdotal reports by examiners suggest a candidate's performance on a case can be evaluated in less than three minutes. In psychiatry the case stimulus is a live patient in one of two stations for one hour of the exam.[5] Exams are administered as quarter or half-day sessions over multiple days for up to five full days. Cases are changed each half-day of administration to reduce risks of cheating.

Standardization of examiners

Certifying Boards use these means to standardize examiners:

1. Select only examiners who are board certified, have appropriate expertise, and are respected in the specialty;
2. Train examiners in how to manage an exam session, evaluate candidates, and record scores;
3. Evaluate examiners on their performance and provide feedback to them.
4. Analyze scoring data for examiner bias and adjust scores for examiner severity.
5. Retain as examiners only those who conform to the board's standards for examination procedures and scoring candidates.

Examiner training varies across exams from a one-hour re-orientation session for returning experienced examiners to six to 10 hours for new and returning examiners. The training involves reviewing the clinical cases and props, videotaped and written instructions for study at home, and role-playing exam simulations in teams. Verbal and written feedback is provided to examiners by senior examiners and in reports comparing scoring patterns and pass/fail rates. Training does make a significant difference as reported by Des Marchais and Jean for exams administered by the Royal College of Physicians and Surgeons of Canada.[6] All of the training efforts are intended to calibrate the examiner's performance to match the boards' standards for questioning and scoring candidates.

Standardization of Cases

An expert committee creates a pool of patient care cases for most exams, except when the candidate's actual patient cases are used in the exam. When the case pool is used a subgroup of examiners selects patient cases from the pool for their stations. When candidates' cases are used candidates supply actual patient records without patient identifiers (two exams), or a brief synopsis of the clinical case including follow-up care and pathology findings (five exams). Four exams use a mix of board and candidate cases, and one uses only candidate cases. When the

candidate's cases are used the oral examination is called a "chart stimulated recall oral exam." Two exams use life patients as the case.

Scoring

In principle, the expectation is to generate as many separable scores as feasible from as many cases as possible. The critical factor in determining scorable points is the number of cases. Using more than one examiner in a session provides a more reliable score for the case but does not increase the number of points for scoring the case. The total number of scorable points varies between eight and 80. A typical examination has 25 to 42 scorable points. Most exams use a rating scale without behavioral anchors. Some use pass/fail points only. A few generate a conditional pass score applying a post hoc scoring rubric to separate passing from failing performance. Scored separately in most exams are these attributes of clinical judgment: use of clinical knowledge, diagnostic decisions, treatment decisions, and management of complex or special problems. Examiners are expected to justify in writing negative decisions. Some boards also include a global pass/fail rating for each case and for each examination session. The passing level is based upon scores of a reference group of examines except for one exam which uses a modified Angoff procedure to set the passing level.

Reliability. In some but not all examinations the cases and examiners are calibrated using item response theory methods (IRT) to adjust for item difficulty and correct for examiner variability.[7] Typical score reliability statistics using IRT methods for two hour exams range between 0.80 and 0.95 (personal communication, M. Lunz). McGuire and colleagues reported score reliabilities of 0.84 in early work on what was then called "role-playing orals," a format that resembles current standardized oral exams.[8] Using generalizability theory a score reliability of 0.60 was reported for a two hour standardized oral examination (Personal Communication, H. Ham).

Discussion

Some psychometricians have questioned the value of the standardized oral examination as an assessment method to assess clinical judgment based upon two arguments:

1. These tests are unreliable given the small sample size of "test items" and low reported "inter-rater" reliability statistics.
2. Why even use performance assessments when the multiple-choice question (MCQ) item format can be used to measure clinical judgment?

In response to the first argument I find it curious that the reliability statistic recommended by these pundits is "inter-rater reliability" rather than obtaining a "score reliability or better named "score reproducability."[9] Inter-rater reliability suggests that the test item is the examiner when in fact the test item is the interaction term of the "case by examiner." Written exams of the same length, two hours, have reliabilities below 0.76 for multiple-choice question item formats.[10] It would appear that the better constructed standardized oral examinations are at least as reliable if not more reliable than the standardized MCQ examinations of similar length.

There is little evidence that an MCQ item format can be used to measure clinical judgment. First, correlations between the two formats are modest to low suggesting they measure different

abilities.[11] Second, when test writers are asked to create MCQs to measure clinical judgment they are rarely successful. An exception is an experiment at the Educational Testing Service to create "open problem space" MCQs for the Graduate Record Exam (GRE).[12] But, these item types are difficult to construct and have not been implemented to my knowledge in the GRE or other standardized exams. Lastly, research by the developers of the "key features" cases found that using lists of responses in place of open-ended options artificially inflated scores of low performers.[13] It seems the MCQ item format even with "key features" cases provides cues that affect measurement accuracy. It is time to dispel the illusion that an MCQ item format is practical to assess reasoning, judgment or practice performance.[14]

In summary the standardized oral examination provides one of two potential means to measure the clinical judgment of professionals. The second option is to limit the measurement to clinical reasoning skills using "key features" cases. Fredricksen and colleagues offer guidance by creating a new measurement paradigm to use in practice performance assessment and tests of higher order thinking.[15] Let's hope others follow their lead.

Contact:
Philip G. Bashook, EdD
American Board of Medical Specialties
1007 Church Street (Suite 404)
Evanston, IL 60201-5913 USA
847-491-9091
pgb@abms.org

References

1. Mancall EL. The oral examination: an historical perspective. In Mancall EL, Bashook, PG (editors). *Assessing clinical reasoning: the oral examination and alternative methods*. Evanston, IL: American Board of Medical Specialties, 1995, pp. 3-7.
2. American Board of Medical Specialties. *Annual Report and Reference Handbook*. Evanston, Illinois: American Board of Medical Specialties. 1999, p. 102, Table 3.
3. Bordage G. Elaborated knowledge: A key to successful diagnostic thinking. *Acad. Med.* 1994; 69:883-885.
4. Halio JL. Ph.D's and the oral examination. *J. Higher Educ.* 1963; 34:140-152.
5. Juul D, Scheiber SC. The Part II psychiatry examination: facts about the oral examination. IN Shore JH, Scheiber SC (Eds). *Certification, recertification, and lifetime learning in psychiatry*. Washington, DC: Am Psychiatr Press, Inc. 1994, and pp. 71-90.
6. Des Marchais JF, Jean P. Effects of examiner training on open-ended, high taxonomic level questioning in oral certification examinations. *Teach & Learn. Med.* 1993: 5:24-28.
7. Lunz ME, Stahl JA, Wright BD. Interjudge reliability and decision reproducibility. *Educ. Psychol. Meas.* 1994; 54:913-925.
8. Levine HG, McGuire CH. The use of role-playing to evaluate affective skills in medicine. *J. Med. Educ.* 1970; 45:700-705.
9. Lunz ME. Statistical methods to improve decision reproducibility. In Mancall EL, Bashook, PG (editors). *Assessing clinical reasoning: the oral examination and alternative methods*. Evanston, IL: American Board of Medical Specialties, 1995, pp. 97-106.
10. Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Med. Educ.* 1985; 19:238-247.
11. Muzzin LJ and Hart L. Oral examinations. In Neufeld VR and Norman GR (editors). *Assessing Clinical Competence*. New York: Springer Publishing Co. 1985, pp 71-93, Table 5.1.
12. Enright MK, Tucker CB, Katz IR. A cognitive analysis of solutions for verbal, informal, and formal-deductive reasoning problems. Princeton, NJ: Educational Testing Service, 1995 (GRE Board report No. 90-04P).
13. Page G, Bordage G. The Medical Council of Canada's key features project: a more valid written examination of clinical decision-making skills. *Acad. Med.* 1995; 70:104-110.
14. Nichols, P, Sugrue B, The lack of fidelity between cognitively complex constructs and conventional test development practices. *Educ. Meas.* 1999; 18-29.
15. Frederiksen N, Mislevy RJ, Bejar I. (Eds). *Test Theory for a new Generation of Tests*. Hillsdale, New Jersey: L. Erlbaum Assoc, 1993.